Data Preparation for Data Science

Data Assembly

Data Quality for Analytics

Feature Generation

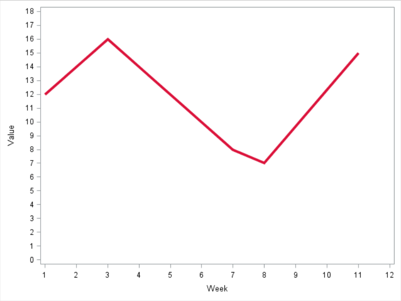
Data Quality for Analytics Using SAS

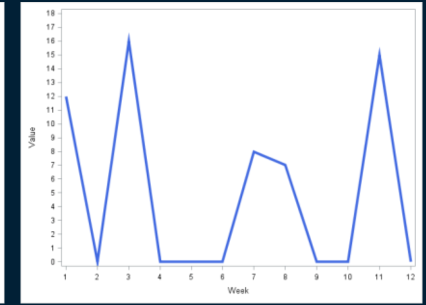
Gerhard Svolba

SAS Press

#3

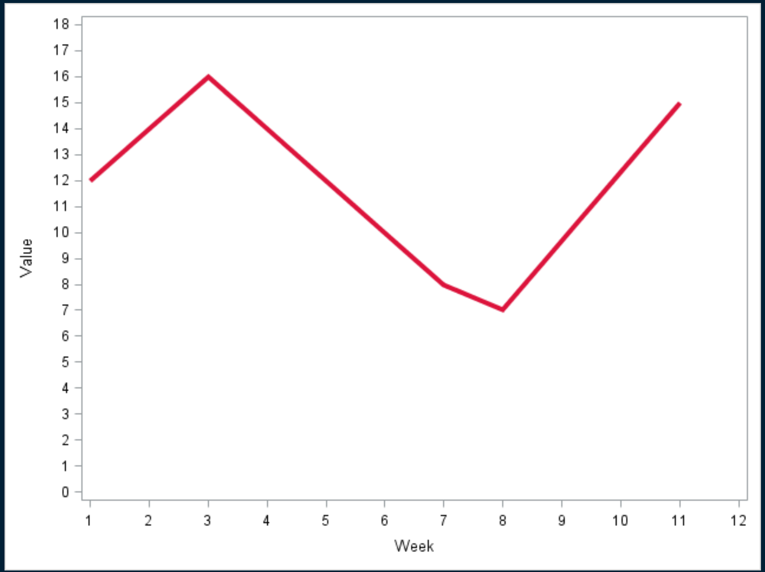
Detecting and Treating Missing Values in Longitudinal data

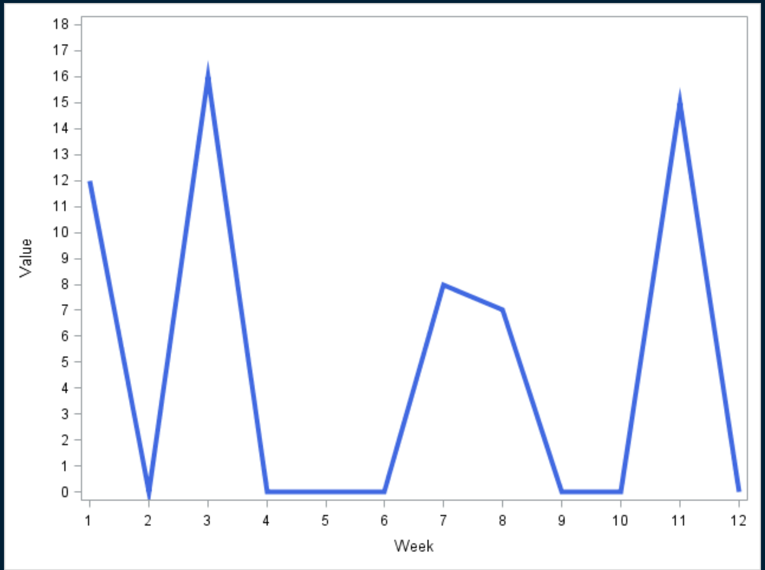




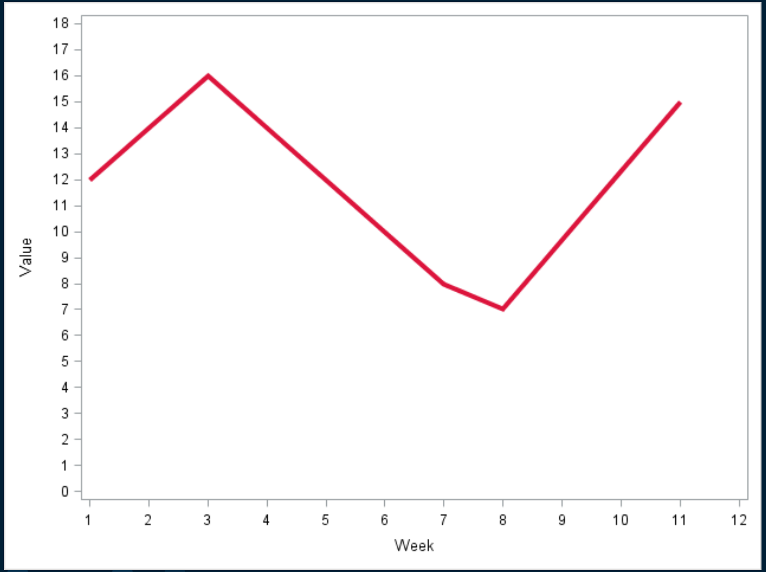
Gerhard Svolba
Data Scientist
SAS Austria

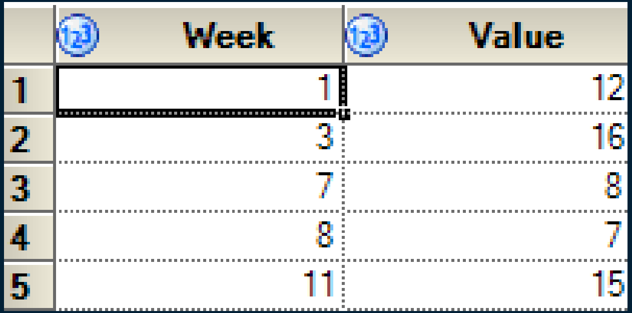
Are these two graphs based on the same data?





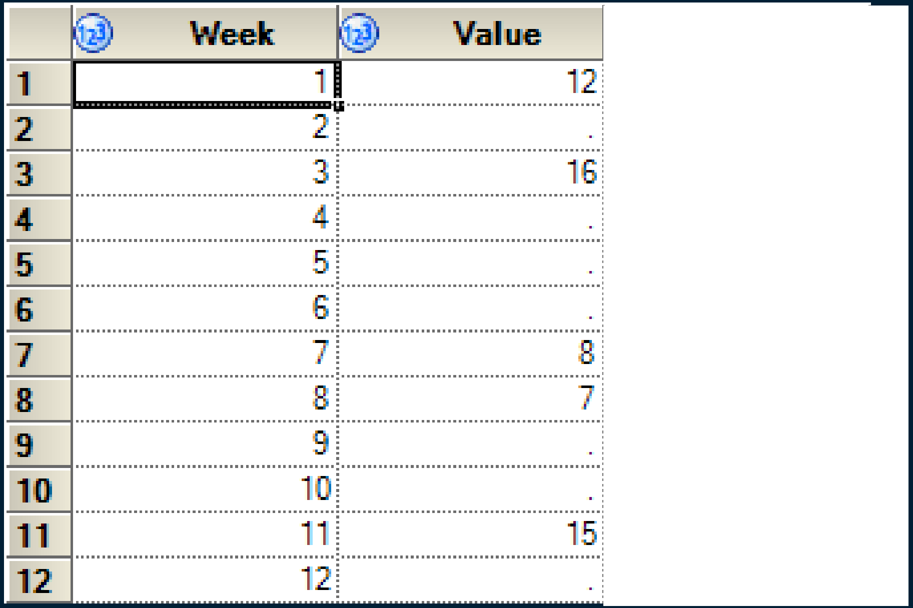
Assuming observations at certain points in time



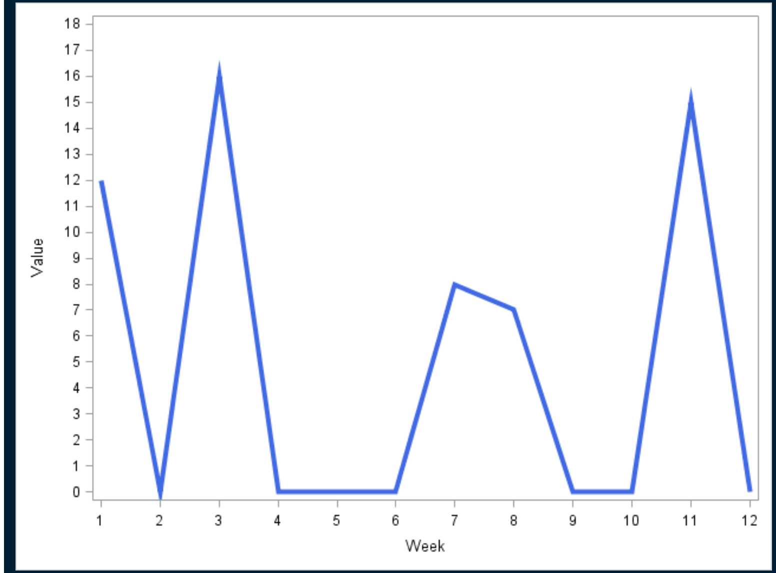


	Week	Value
1	1	12
2	3	16
3	7	8
4	8	7
5	11	15

Assuming event being accumulated per time period
No event – no record in the analysis data

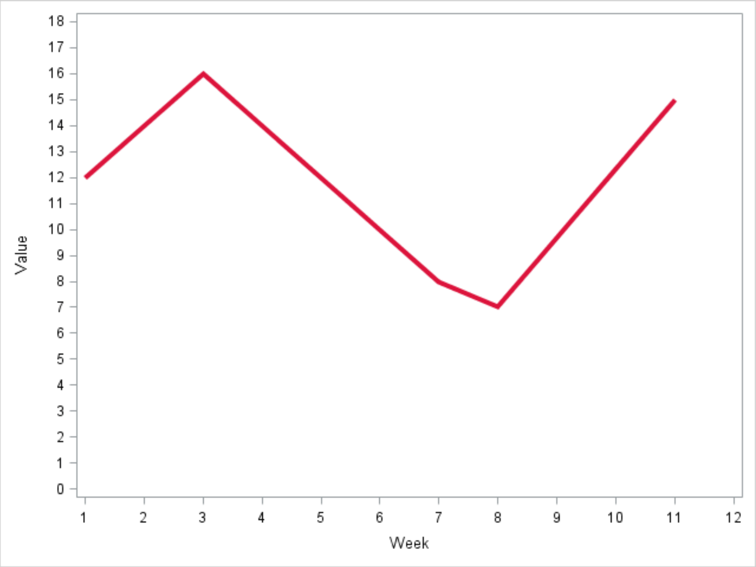


	Week	Value
1	1	12
2	2	.
3	3	16
4	4	.
5	5	.
6	6	.
7	7	8
8	8	7
9	9	.
10	10	.
11	11	15
12	12	.

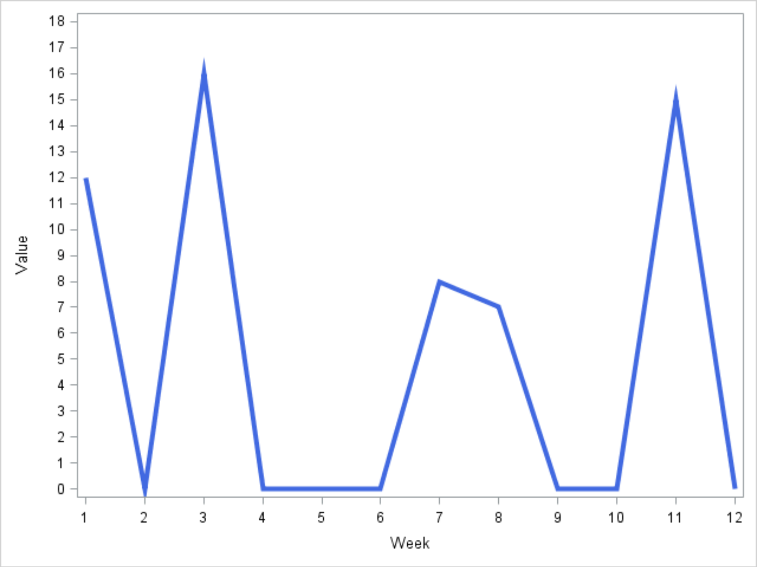


Which is the correct representation?

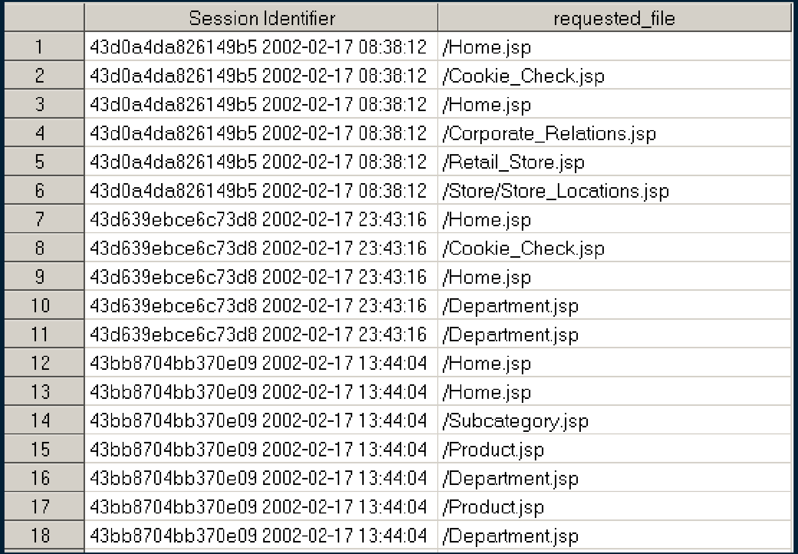
Outside Temperature (°C)



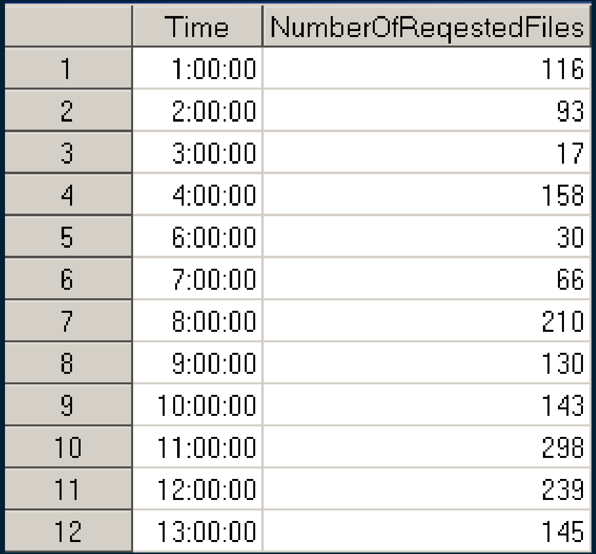
Precipitation (mm)



Transactional Data or Timeseries Data?

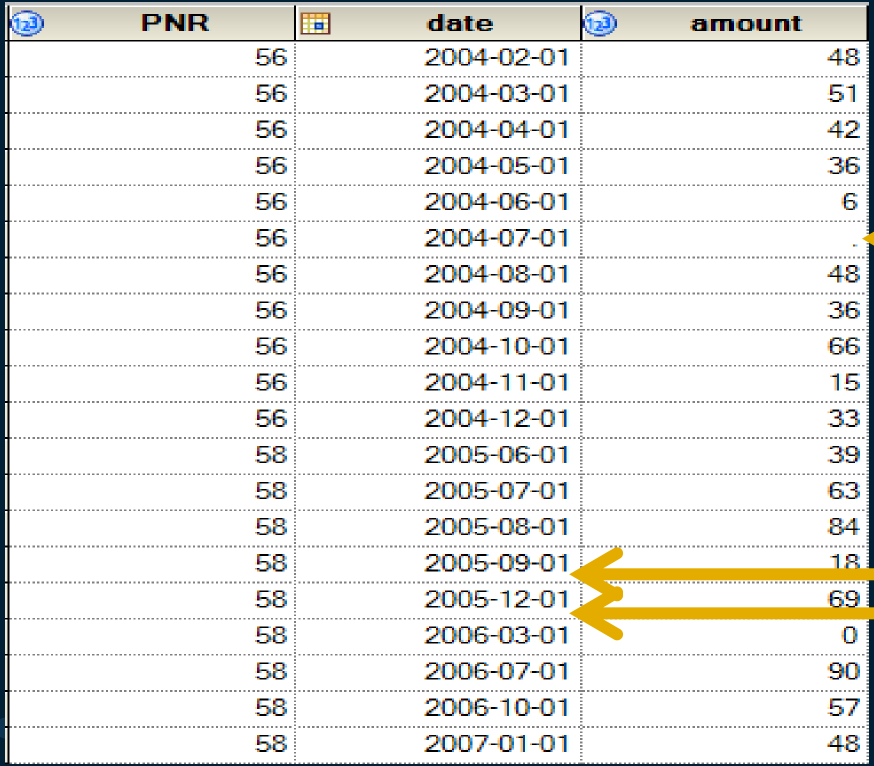


	Session Identifier	requested_file
1	43d0a4da826149b5 2002-02-17 08:38:12	/Home.jsp
2	43d0a4da826149b5 2002-02-17 08:38:12	/Cookie_Check.jsp
3	43d0a4da826149b5 2002-02-17 08:38:12	/Home.jsp
4	43d0a4da826149b5 2002-02-17 08:38:12	/Corporate_Relations.jsp
5	43d0a4da826149b5 2002-02-17 08:38:12	/Retail_Store.jsp
6	43d0a4da826149b5 2002-02-17 08:38:12	/Store/Store_Locations.jsp
7	43d639ebce6c73d8 2002-02-17 23:43:16	/Home.jsp
8	43d639ebce6c73d8 2002-02-17 23:43:16	/Cookie_Check.jsp
9	43d639ebce6c73d8 2002-02-17 23:43:16	/Home.jsp
10	43d639ebce6c73d8 2002-02-17 23:43:16	/Department.jsp
11	43d639ebce6c73d8 2002-02-17 23:43:16	/Department.jsp
12	43bb8704bb370e09 2002-02-17 13:44:04	/Home.jsp
13	43bb8704bb370e09 2002-02-17 13:44:04	/Home.jsp
14	43bb8704bb370e09 2002-02-17 13:44:04	/Subcategory.jsp
15	43bb8704bb370e09 2002-02-17 13:44:04	/Product.jsp
16	43bb8704bb370e09 2002-02-17 13:44:04	/Department.jsp
17	43bb8704bb370e09 2002-02-17 13:44:04	/Product.jsp
18	43bb8704bb370e09 2002-02-17 13:44:04	/Department.jsp



	Time	NumberOfReqestedFiles
1	1:00:00	116
2	2:00:00	93
3	3:00:00	17
4	4:00:00	158
5	6:00:00	30
6	7:00:00	66
7	8:00:00	210
8	9:00:00	130
9	10:00:00	143
10	11:00:00	298
11	12:00:00	239
12	13:00:00	145

Differentiate between explicit and implicit missing values in longitudinal data

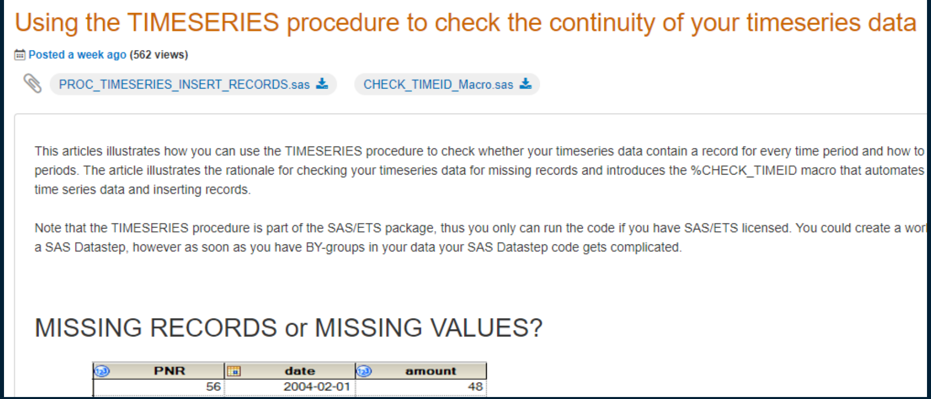


PNR	date	amount
56	2004-02-01	48
56	2004-03-01	51
56	2004-04-01	42
56	2004-05-01	36
56	2004-06-01	6
56	2004-07-01	.
56	2004-08-01	48
56	2004-09-01	36
56	2004-10-01	66
56	2004-11-01	15
56	2004-12-01	33
58	2005-06-01	39
58	2005-07-01	63
58	2005-08-01	84
58	2005-09-01	18
58	2005-12-01	69
58	2006-03-01	0
58	2006-07-01	90
58	2006-10-01	57
58	2007-01-01	48

Existing Record
Value Missing

Missing Record
No Continuity

Two related Articles at Communities.sas.com



Using the TIMESERIES procedure to check the continuity of your timeseries data

Posted a week ago (562 views)

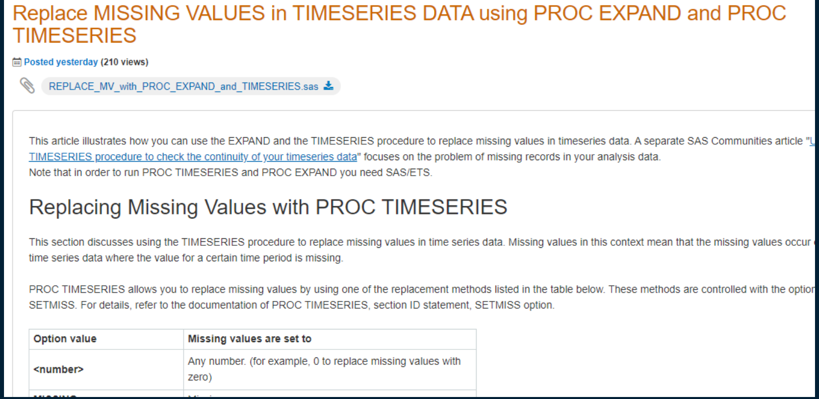
PROC_TIMESERIES_INSERT_RECORDS.sas CHECK_TIMEID_Macro.sas

This articles illustrates how you can use the TIMESERIES procedure to check whether your timeseries data contain a record for every time period and how to
periods. The article illustrates the rationale for checking your timeseries data for missing records and introduces the %CHECK_TIMEID macro that automates
time series data and inserting records.

Note that the TIMESERIES procedure is part of the SAS/ETS package, thus you only can run the code if you have SAS/ETS licensed. You could create a wor
a SAS Datastep, however as soon as you have BY-groups in your data your SAS Datastep code gets complicated.

MISSING RECORDS or MISSING VALUES?

PNR	date	amount
56	2004-02-01	48



Replace MISSING VALUES in TIMESERIES DATA using PROC EXPAND and PROC TIMESERIES

Posted yesterday (210 views)

REPLACE_MV_with_PROC_EXPAND_and_TIMESERIES.sas

This article illustrates how you can use the EXPAND and the TIMESERIES procedure to replace missing values in timeseries data. A separate SAS Communities article "U
TIMESERIES procedure to check the continuity of your timeseries data" focuses on the problem of missing records in your analysis data.
Note that in order to run PROC TIMESERIES and PROC EXPAND you need SAS/ETS.

Replacing Missing Values with PROC TIMESERIES

This section discusses using the TIMESERIES procedure to replace missing values in time series data. Missing values in this context mean that the missing values occur
time series data where the value for a certain time period is missing.

PROC TIMESERIES allows you to replace missing values by using one of the replacement methods listed in the table below. These methods are controlled with the optior
SETMISS. For details, refer to the documentation of PROC TIMESERIES, section ID statement, SETMISS option.

Option value	Missing values are set to
<number>	Any number. (for example, 0 to replace missing values with zero)

Links and Papers

- Using the TIMESERIES procedure to check the continuity of your timeseries data
- Replace MISSING VALUES in TIMESERIES DATA using PROC EXPAND and PROC TIMESERIES
- SGF-Paper: Want an Early Picture of the Data Quality Status of Your Analysis Data? SAS® Visual Analytics Shows You How

Replacing and interpolating missing values in longitudinal data with SAS

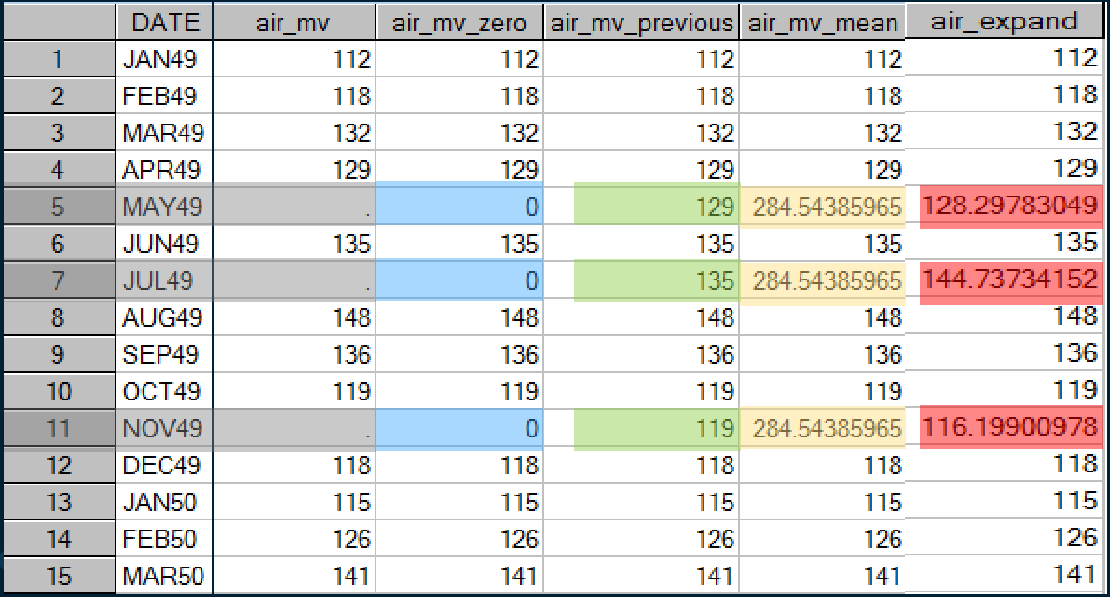
Insert missing records

Replace with 0

Replace with last known value

Replace with mean

Interpolate based on splines



	DATE	air_mv	air_mv_zero	air_mv_previous	air_mv_mean	air_expand
1	JAN49	112	112	112	112	112
2	FEB49	118	118	118	118	118
3	MAR49	132	132	132	132	132
4	APR49	129	129	129	129	129
5	MAY49	.	0	129	284.54385965	128.29783049
6	JUN49	135	135	135	135	135
7	JUL49	.	0	135	284.54385965	144.73734152
8	AUG49	148	148	148	148	148
9	SEP49	136	136	136	136	136
10	OCT49	119	119	119	119	119
11	NOV49	.	0	119	284.54385965	116.19900978
12	DEC49	118	118	118	118	118
13	JAN50	115	115	115	115	115
14	FEB50	126	126	126	126	126
15	MAR50	141	141	141	141	141

PROC TIMESERIES and PROC EXPAND in SAS Viya fulfill these tasks!

Missing Values in TimeSeries Data - Get into the details!

- Missing Value or Missing Record?
- What should I impute:
  0 or an interpolated value?
- Visualize the completeness structure of your time series data
- Encourage „Analytic Awareness“ in Data Prepration and Data Quality

Data Quality for Analytics Using SAS®

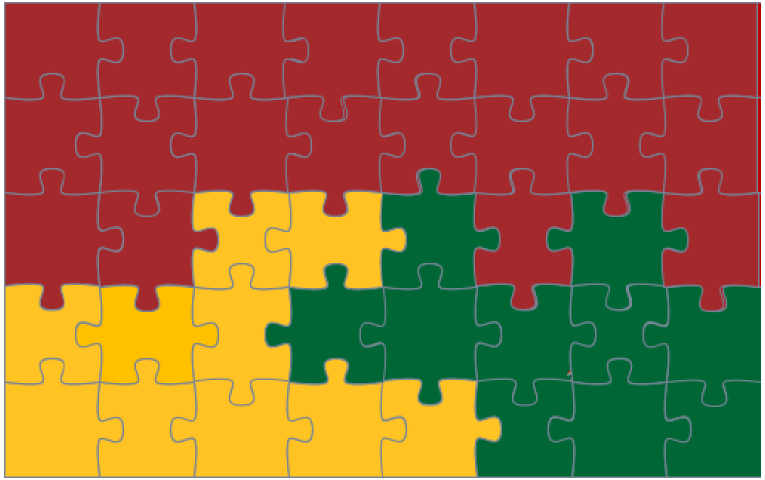
Dr. Gerhard Svolba

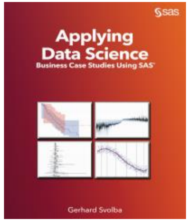
SAS Press

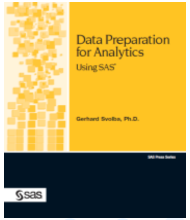
Data Preparation for Data Science

- Data Assembly
- Data Quality for Analytics
- Feature Generation

Gerhard Svolba,
Data Scientist @SAS
mailto:sastools.by.gerhard@gmx.net







Articles and Blogs Medium LinkedIn

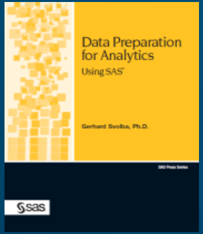
Webinars YouTube

Tipps & Tricks sas COMMUNITIES

Macros & Downloads

Get access to more content:

SAS DACH @Youtube:	https://www.youtube.com/user/SASsoftwareGermany
Blogs on LinkedIn:	https://www.linkedin.com/in/gerhardsvolba/
Twitter:	https://twitter.com/gsvolba
Content on Github:	https://github.com/gerhard1050
Books @SAS-Press:	https://support.sas.com/svolba



Data Preparation for Analytics Using SAS

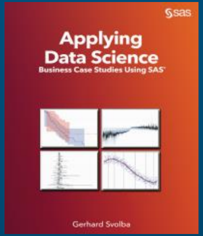
Gerhard Svolba, Ph.D.

SAS Press Series

Data Quality for Analytics Using SAS

Dr. Gerhard Svolba

SAS Press



Applying Data Science

Business Case Studies Using SAS

Gerhard Svolba